Critical appraisal of health literacy indices revealed variable underlying constructs, narrow content and psychometric weaknesses

Joanne E. Jordan a,*, Richard H. Osborne a,b, Rachelle Buchbinder c,d

a Department of Medicine, Centre for Rheumatic Diseases, The Royal Melbourne Hospital, The University of Melbourne, Parkville, Victoria, Australia
b Public Health Research and Evaluation and Policy Cluster, School of Health and Social Development, Deakin University, Melbourne, Victoria, Australia
c Monash Department of Clinical Epidemiology at Cabrini Hospital, Monash University, Victoria, Australia
d Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Cabrini Medical Centre, Victoria, Australia

Accepted 13 April 2010

Abstract

Objective: Health literacy refers to an individual’s ability to seek, understand, and use health information. A range of indices exist that purport to measure health literacy across individuals and populations. This study aimed to review the development and content of existing indices and to critically appraise their properties.

Study Design and Setting: Using standardized search terms, published generic health literacy indices (1990–2008) were identified. Using a methodological framework, each was evaluated for purpose, validity (face, content, construct), reliability, responsiveness, feasibility, and generalizability.

Results: Nineteen instruments were evaluated. Three measurement approaches were identified: direct testing of individual abilities, self-report of abilities, and population-based proxy measures. Composition of underlying constructs and content varied widely across instruments, and none appeared to fully measure a person’s ability to seek, understand, and use health information. The content was focused primarily on reading comprehension and numeracy; scoring categories were poorly defined and may not be mutually exclusive, and few indices had been assessed for reliability.

Conclusion: Health literacy is not consistently measured, making it difficult to interpret and compare health literacy at individual and population levels. Empirical evidence demonstrating validity and reliability of existing indices is required, and more comprehensive health literacy instruments need to be developed.

Keywords: Health literacy; Tools; Assessments; Psychometrics; Validity; Reliability

1. Introduction

As a result of health care changes toward a more patient-centered care approach [1,2], there is an expectation that patients take an increased role in decisions about their health. For patients to function effectively within this shifting paradigm, they require a basic set of skills to seek, understand, and use health information, a concept referred to as “health literacy” [3,4]. Suboptimal health literacy skills reduce the likelihood of maintaining good health, independent of other sociodemographic factors [5,6], and are associated with increased health care costs [7].

A range of indices have been developed to measure health literacy [8,9]. To provide credible information to inform clinical practice and health policies and programs, it is imperative that indices have sound psychometric properties. These include the following: validity—the extent to which an instrument measures what it purports to measure when properly administered [10]; reliability—the extent to which the obtained scores are free from measurement error [11]; feasibility—ease of administration; generalizability—use across and within fields; and in some circumstance, responsiveness—the ability of indices to detect change [12]. The purpose of this work was to review the literature on existing health literacy indices and perform a critical appraisal of their concept, content, and psychometric properties.

2. Method

Medline, PubMed, and PsychInfo databases were searched for publications of generic health literacy indices
What is new?

Key finding
A key finding of this critical appraisal was that limited empirical evidence exists on the reliability and construct validity of health literacy measures. This raises uncertainty about the accuracy of data being produced in relation to health literacy levels at an individual and population level. Furthermore, this research demonstrates that great variation exists across indices in terms of content across domains, methods of item selection and scoring and how individuals are categorised or classified in terms of health literacy levels. The variation and weak psychometric data makes it difficult to compare or pool results across studies and hinders the establishment of clear benchmarks for policy and program development aimed at addressing suboptimal health literacy.

What this adds to what is known?
This appraisal, using a structured methodological framework, provides new and synthesised information for researchers and clinicians on the strengths and limitations of current indices of health literacy. Although previous literature acknowledges the narrow approach taken to the measurement of health literacy, the systematic presentation of the validity, reliability and applicability of these instruments provides more detail relating to psychometric properties of existing instruments.

What is the implication, what should change now?
This review will support clinicians, researchers and policymakers to qualify the findings and recommendations derived from the use of health literacy indices. Further research is required to obtain empirical evidence across different populations and settings of the construct validity and reliability of existing measures. New measures which incorporate broader constructs of health literacy are required and are needed to advance this field.

between 1990 and 2008 inclusive, using the following search terms: health, literacy, health literacy, instrument, tool, and assessment. The search was limited to publications in English language. Instruments that were not generic, that is, specific to particular fields or groups, were excluded. Additional instruments were identified through manual searching of the references of relevant published studies and by consulting experts in the field.

All health literacy indices were evaluated for purpose, face validity, content validity, construct validity, reliability, responsiveness, feasibility, and generalizability, using a methodological framework developed for the evaluation of health assessment indices [13]. Categories within the framework were adapted where necessary to make the appraisal relevant for health literacy indices (Tables 1 and 2). Two raters (J.E.J. and R.B.) critically appraised all instruments independently, and disagreements were resolved by consensus or discussion with the third author (R.H.O.).

3. Results

3.1. Questionnaires

Nineteen health literacy instruments met the prespecified selection criteria and were included in the appraisal (Fig. 1) (Tables 1, A1, and A2). There were 12 original instruments [8,9,14–23] and seven derivatives that were short-form versions or adaptations of original instruments [24–30]. There were also Spanish versions of two of the instruments [9,16] and a Chinese version of a short-form instrument [31]. As these were direct translations of original instruments, they were excluded.

There were three main approaches for measuring health literacy: direct testing of an individual’s abilities (Table 1), elicitation of self-report of abilities (Table A1 [available on the journal’s Web site at www.jclinepi.com]), and proxy measures of health literacy in the population (Table A2 [available on the journal’s Web site at www.jclinepi.com]).

3.1.1. Instruments that directly test an individual’s abilities

3.1.1.1. Rapid Estimate of Adult Literacy in Medicine and derivatives. The Rapid Estimate of Adult Literacy in Medicine (REALM) was developed in the United States to assist physicians to identify adults with limited reading skills in the primary care setting (Table 1) [8]. It was modeled on the Wide Range Achievement Test (WRAT), a standardized literacy test that assesses the pronunciation of a list of words that the respondent reads aloud [8]. The 125 words, chosen from materials used in primary care, are arranged in three columns according to the number of syllables in ascending order of difficulty. A point is allocated for each correctly pronounced word. A shortened version, the REALM-S, comprising 66 items, was developed to reduce administration time and enhance uptake within the clinical setting [24]. It takes 1–2 minutes to complete. The raw score for the REALM-S (0–66) is converted to a US school grade estimate of reading ability (Table 1).

Other derivatives include the REALM-shortened version (REALM-R) (eight items) [26] and REALM-Short Form (REALM-SF) (seven items) [27]. The Rapid Estimate of Adolescent Literacy in Medicine (REALM-Teen), another derivative, is tailored to screen youth in grade 6–12 (ages 10–19 years) for below-grade reading in the health care setting [25]. Words for the REALM-Teen were selected
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Country of origin</td>
<td>USA</td>
<td>USA</td>
<td>USA</td>
<td>USA</td>
<td>USA</td>
<td>USA</td>
<td>USA</td>
<td>USA</td>
<td>USA</td>
</tr>
<tr>
<td>Stated expertise of developers</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Panel of doctors, nurses, social workers, psychologists, and educators</td>
<td>Experience working with Spanish-speaking patients in educational and medical settings</td>
<td>Not specified</td>
<td>Literacy expert</td>
<td>Not stated</td>
<td>Public health experts</td>
<td>Panel of health literacy experts</td>
</tr>
</tbody>
</table>

1. Purpose

| a. Stated purpose and population | Identify patients with limited reading skills and estimate patient reading levels in primary care settings | Identify patients with low reading levels in primary care settings | Screen youth in grades 6–12 for below-grade reading | Develop an easy-to-use health literacy test for Spanish speakers in health care settings | To identify illiterate patients of high school age or older in the general community | Understand and measure functional health literacy in patients in health care settings | Measure patients’ ability to read and understand health-related materials in the health care system | Assess health literacy in Hebrew patients in Israeli health care system | Screen for limited literacy in patients in primary health care settings |

2. Method of development

| a. How was the instrument developed (or shortened if more applicable) | Words in instrument chosen from patient education materials and patient intake forms used in university-based primary care clinics. Method of selection not stated. | Shortened using psychometric estimate of item difficulty and discrimination and frequency of retained words in written material given to patients | 116 words were selected from American Academy of Paediatrics’ adolescent patient education pamphlets. The list was piloted tested on 200 students in grades 6–12 and retention of items based on psychometric estimate of item difficulty, item discrimination, and the panel’s judgement. | Items from REALM translated to Spanish. For the additional comprehension section, words were selected by an expert panel based on fluency in English and Spanish using the Delphi process. | Based on the Wide Range Achievement Test. Words selected from prescription labels and from a medical dictionary. The 42 words were chosen to have a corresponding level of difficulty based on words contained in the WRAT reading test. | Developed from sample of hospital texts, including education materials, diagnostic tests, prescription labels, registration forms by literacy expert or the TOFHLA. | Item selection based on previous study that used the TOFHLA. Three numeracy items directly translated from S-TOFHLA. One item changed to resemble Israeli style of arranging appointments. Reading comprehension passages developed by authors. | Five health-related scenarios were developed by a panel of health literacy experts. Scenarios refined through consultation and scenario selected that was found to have strongest psychometric properties. |
3. Description of instrument
   a. Domain
      - Single domain—125 items.
      - Single domain—66 items.
      - Single domain—66 items.
      - Single domain—50 items.
      - Single domain—42 items.
      - Two domains: (1) reading comprehension—50 items; (2) numeracy—17 items.
      - Two domains: (1) reading comprehension—36 items; (2) numeracy—4 items.
      - Two domains: (1) reading comprehension—8 items; (2) numeracy—4 items.
      - Single domain—6 items.

4. How is it administered
   - Interview administered
   - Interview administered
   - Interview administered
   - Interview administered
   - Numeracy section interview administered
   - Numeracy section interview administered
   - Interview administered

5. Special requirements for administration
   - Knowledge of correct pronunciation of words
   - Knowledge of correct pronunciation of words
   - Knowledge of correct pronunciation of words and provide cue cards
   - Provide cue cards and verbally administer numeracy questions
   - None specified
   - Verbally administer questions

6. Estimated time duration
   - 3–5min
   - 1–2min
   - 2–3min
   - 3–6min
   - 3–5min
   - Upto 22min
   - Less than 10min
   - Not stated
   - Average 2.9min

7. Scoring
   a. How is it scored
      - Single score: 0–125
      - Single score: 0–66
      - Single score: 0–66
      - Single score: 0–50
      - Single score: 0–42
      - Combined weighted score: 0–100
      - Combined weighted score: 0–100
      - Combined score from 0 to 12
      - Single score: 0–6
   b. Scoring categories
      - School grade estimate:
        - 0–78 below third grade; 79–103 for fourth to sixth grade; 104–114 for seventh to eighth grade; 115+ for high school
        - School grade estimate:
        - 0–18 for third grade and below; 19–44 for fourth to sixth grade; 45–60 for seventh to eighth grade; 61–66 for ninth grade and higher
        - School grade estimate:
        - 0–37 below third grade; 38–47 for fourth to fifth grade; 48–58 for sixth to seventh grade; 59–62 for eighth to ninth grade; 63–66 for tenth grade and higher
        - School grade levels but categories are not specified
        - 2 categories—<37: inadequate health literacy; 38–50: adequate health literacy
        - School grade levels but categories are not specified
        - 0–59: inadequate health literacy; 60–74: marginal health literacy; 75–100 adequate health literacy

Abbreviations: REALM, Rapid Estimate of Adult Literacy in Medicine; REALM-S, Rapid Estimate of Adult Literacy in Medicine shortened version; REALM-Teen, Rapid Estimate of Adolescent Literacy in Medicine; SAHLSA, Short Assessment of Health Literacy for Spanish-speaking Adults; MART, Medical Achievement Reading Test; TOFHLA, Test of Functional Health Literacy in Adults; S-TOFHLA, Test of Functional Health Literacy in Adults short version; HHLT, Hebrew Health Literacy Test; NVS, Newest Vital Sign.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Is the instrument based on an underlying conceptual framework?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Not specified</td>
</tr>
<tr>
<td>3. Content validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Were intended domains (i.e., relevant areas to be included and excluded) clearly stated?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>b. Are all relevant components of each domain included?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>4. Face validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. On the face of it does it describe the intended purpose</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>b. Are the definitions of each category clearly specified?</td>
<td>Partial</td>
<td>Partial</td>
<td>Partial</td>
<td>No</td>
</tr>
<tr>
<td>c. Are these definitions acceptable?</td>
<td>Partial—provides a school grade estimate for reading ability with descriptions for whether individual may/may not need low literacy materials</td>
<td>Partial—no clear differentiation between 4–6 and 7–8 grade range estimates in terms of what materials will be suitable for</td>
<td>Partial—no accompanying definitions with school grade range estimates</td>
<td>No</td>
</tr>
<tr>
<td>d. Are the methods for determining the presence and/or absence of criteria described and acceptable?</td>
<td>No. Criteria not described and correct pronunciation of words will vary amongst population groups</td>
<td>Partial, dictionary pronunciation is the scoring standard for ruling whether words are pronounced correctly</td>
<td>Partial, dictionary pronunciation is the basis of scoring standard</td>
<td>Partial, didn’t fully detail how different idiomatic expressions across Spanish-speaking countries were adapted</td>
</tr>
<tr>
<td>e. Do the scoring categories sufficiently discriminate considering the stated purpose?</td>
<td>Partial—hard to distinguish between grade equivalents 4–6 and 7–8 in terms of what health information materials would be used</td>
<td>Partial. The categories do not provide clear directions in terms of which groups categorically will need low literacy materials</td>
<td>Partial. Provides a school grade estimate but there is no corresponding description to assist what materials an individual may need.</td>
<td>Partial—identifies between inadequate and adequate health literacy but no accompanying descriptions</td>
</tr>
<tr>
<td>5. Construct validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Does the instrument perform in expected ways when compared with other health literacy/literacy indices?</td>
<td>Yes, high correlation with standardized general reading tests: (1) SORT-R—Pearson correlation coefficient $r = 0.95$; (2) PIAT-R—Pearson correlation coefficient $r = 0.94$</td>
<td>Yes, high correlation with standardized general reading tests: (1) SORT-R—Pearson correlation $r = 0.96$; (2) PIAT-R—Pearson correlation $r = 0.97$, and (3) WRAT—Pearson correlation $r = 0.88$. Moderately high correlation with (1) TOFHLA—Spearman rank correlation $r = 0.74$; (2) Basic Skills Assessment Initial Test (UK)—Pearson correlation coefficient $r = 0.70$; (3) REALM-R—Spearman rank correlation $r = 0.72$. Poor correlation with (1) TORCH (Aus)—Pearson correlation coefficient $r = 0.36$</td>
<td>Yes high correlation with (1) WRAT-3—Pearson correlation $r = 0.83$ and (2) SORT-R—Pearson correlation $r = 0.93$</td>
<td>Partial, correlated with TOFHLA (Spanish version)—Pearson correlation $r = 0.65$. Difference expected given no numeracy component for SAHLSA. Confirmatory factor analysis model: RMSEA = 0.04 and CFI = 0.85</td>
</tr>
<tr>
<td>6. Has sensitivity to change been demonstrated?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>7. Reliability</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Has the reliability been measured?</td>
<td>Yes, test-retest reliability coefficient 0.98</td>
<td>Yes, test-retest reliability coefficient 0.99</td>
<td>Yes, test-retest coefficient 0.98</td>
<td>Yes, test-retest reliability coefficient 0.86</td>
</tr>
<tr>
<td>b. Interrater reliability</td>
<td>0.99</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Not stated</td>
</tr>
<tr>
<td>c. Intrarater reliability</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Yes, Coefficient alpha 0.94</td>
</tr>
<tr>
<td>d. Internal consistency</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Not stated</td>
</tr>
<tr>
<td>7. Feasibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Is it simple to understand?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>b. Is it easy to perform and administer?</td>
<td>Partial, knowledge of correct pronunciation of words required.</td>
<td>Partial, knowledge of correct pronunciation of words required.</td>
<td>Partial, knowledge of correct pronunciation of words required.</td>
<td>Partial, uncertainty in differentiating between different Spanish expressions</td>
</tr>
</tbody>
</table>

Abbreviations: SORT-R, Slosson Oral Reading Test; PIAT-R, Peabody Individual Achievement Test Revised; WRAT, Wide Range Achievement Test; TORCH, Test of Reading Comprehension; RMSEA, root mean square error of approximation; CFI, comparative fit index; HHLT, Hebrew Health Literacy Test.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Functional health literacy, defined as reading, writing, and numeracy skills</td>
<td>Functional health literacy</td>
<td>Functional health literacy</td>
<td>Not specified</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Partial, justification for inclusion of numeracy</td>
</tr>
<tr>
<td>Partial, illiteracy defined as unable to read and write but only look at reading ability.</td>
<td>Partial, writing ability (stated as component of functional health literacy) not tested</td>
<td>Partial, does not include testing of writing abilities</td>
<td>Partial, writing ability not tested.</td>
<td>Partial—overemphasis on numeracy items.</td>
</tr>
<tr>
<td>Partial</td>
<td>Partial</td>
<td>Partial</td>
<td>Partial</td>
<td>Partial</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Partial, correct word pronunciation not specified</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Unable to be evaluated. Based on WRAT scoring categories but scoring levels and categories not outlined</td>
<td>Partial—categorizes into inadequate, marginal, and adequate functional health literacy but no generalizable descriptions</td>
<td>Partial—categorize into inadequate, marginal, and adequate functional health literacy but no generalizable descriptions</td>
<td>Partial—categories differ from that of S-TOFHLA and use low/marginal/high health literacy categories but no accompanying definitions</td>
<td>Partial—there seems to be a bit of overlap for scores 0–1 and 2–3 in terms of both having descriptions of inadequate/marginal literacy. No definitions provided</td>
</tr>
<tr>
<td>No. No comparisons conducted.</td>
<td>Partial. Correlated with (1) WRAT—Spearman rank correlation $r = 0.74$ and (2) REALM—Spearman rank correlation $r = 0.84$. The latter correlation is possibly higher than expected given that the TOFHLA is measuring comprehension and numeracy as well as reading ability. Poor correlation with (1) TORCH (Aus)—Pearson correlation coefficient $r = 0.39$ and (2) NVS Spearman rank correlation $r = 0.49$</td>
<td>Yes, reading comprehension section correlates highly with REALM—Spearman correlation coefficient $r = 0.81$ although low correlation for the numeracy section with REALM—Spearman correlation coefficient $r = 0.61$ which is to be expected as measuring different constructs</td>
<td>Partial. When results of IHLT compared with proxy measure of patient self-reported reading ability. Pearson correlation $r = 0.67$</td>
<td>No, low correlation with (1) TOFHLA that measures similar constructs—Pearson correlation $r = 0.59$; (2) S-TOFHLA Pearson correlation coefficient $r = 0.6$; (3) REALM Pearson correlation coefficient $r = 0.41$ (primary care clinic) and Spearman rank correlation $r = 0.24$ (Australian population study).</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Not stated</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Not stated</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Yes, Coefficient alpha 0.98</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Yes, Coefficient alpha 0.98</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Yes, Coefficient alpha 0.68 (numeracy) and 0.97 (reading comprehension)</td>
<td>Yes, Coefficient alpha 0.98</td>
</tr>
<tr>
<td>Yes</td>
<td>Partial (length)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

J.E. Jordan et al. / Journal of Clinical Epidemiology 64 (2011) 366–379
by an expert panel from adolescent patient education pamphlets, and it has more school grade categories than the REALM (Table 1).

3.1.1.2. Short Assessment of Health Literacy for Spanish-speaking Adults. The Short Assessment of Health Literacy for Spanish-speaking Adults (SAHLSA) was based on the REALM-S, but it also includes a comprehension test (Table 1) [15]. Respondents read aloud a list of 50 medical terms and also have to nominate a word of similar meaning from a choice of two other words. A point is allocated for each correct answer, and respondents are classified as having either inadequate or adequate health literacy [15].

3.1.1.3. The Medical Achievement Reading Test. Also modeled on the WRAT, the Medical Achievement Reading Test (MART) was designed to identify illiterate patients in a nonthreatening manner [14]. It consists of 42 words, selected from prescription labels and medical dictionary and chosen to reflect the same difficulty level as that of the items in the WRAT. For respondents to feel less intimidated when undergoing testing, the MART uses small print and a glossy cover that creates a glare and makes reading the words difficult to provide individuals with excuses for not being able to read the words. Similar to the WRAT, raw scores are converted to school grade levels; however, scoring and grade-level placement were not provided in the published literature [14, 32].

3.1.1.4. Test of Functional Health Literacy in Adults and derivatives. The Test of Functional Health Literacy in Adults (TOFHLA) was developed in the United States to measure “functional health literacy,” defined as assessing reading, writing, and numeracy skills, in relation to health care (Table 1) [9]. Items were chosen from hospital texts by a literacy expert. A self-administered reading comprehension component includes three passages of texts and contains 50 items. It uses the modified Cloze procedure, where every fifth to seventh word in a passage is omitted, and the respondent selects a response from four options [33]. An interviewer-administered numeracy component includes 17 items, where individuals are presented with cues and respond to questions. Scores for the numeracy domain are transformed to a score out of 50 and added to the score from the reading comprehension domain (total: 0–100).

The raw score for the TOFHLA is converted to one of three categories: inadequate, marginal, or adequate health literacy. These categories were derived from an interpretation of the scores of 2,659 predominantly indigent and minority patients presenting for acute care at two urban hospitals in Atlanta and California, although the method is not explained [34].

A short version, S-TOFHLA, includes 36 reading comprehension and four numeracy items and are also converted to three categories although the score cutoffs are different (Table 1) [30]. Another derivative instrument, the Hebrew Health Literacy Test (HHLT) [29], retains the same structure as the S-TOHFLA, but the items have been adapted to be relevant to the Israeli health care system and is administered in Hebrew. The reading comprehension component has eight items, and the numeracy component has four items. Individuals are classified as having either low, marginal, or high health literacy.

3.1.1.5. Newest Vital Sign. The Newest Vital Sign (NVS) was developed by a panel of health literacy experts to screen for limited literacy in the primary health care setting [16]. It tests reading, interpretation, and numeracy skills. An interviewer asks six questions relating to information contained in a nutritional label from an ice cream container. A point is given for each correct answer. Scores are categorized as high likelihood of marginal/inadequate literacy, possibility of marginal/inadequate literacy, and adequate literacy (Table 1).

3.1.2. Questionnaires that elicit self-report of abilities 3.1.2.1. Set of brief screening questions and derivative. The set of brief screening questions (SBSQ) was developed to detect inadequate or marginal health literacy in a clinical setting (Table A1 [available on the journal’s Web site at www.jclinepi.com]) [17]. The content of the 16 questions was based on five domains identified in a qualitative study of patients with limited health literacy [35]. Respondents are asked to report how frequently they experience problems reading or understanding health information or how confident they are performing these tasks using a 5-point Likert scale (Table A1 [available on the journal’s Web site at www.jclinepi.com]). Based on the comparison with the S-TOFHLA, three questions were selected for the final instrument based on their greater sensitivity and specificity for detecting individuals with inadequate health literacy (Table A3).

A derivative of the SBSQ, the Single-Item Literacy Scale (SILS) asks respondents to report their need for help to read or understand printed health information [28] (Table A1 [available on the journal’s Web site at www.jclinepi.com]). A cutoff score ≥2 indicates difficulty with reading printed health-related information [28].

3.1.2.2. Scales for measuring functional, communicative, and critical health literacy. Developed in Japan by clinicians and researchers, the functional, communicative, and critical health literacy (FCCHL) measures three different levels of health literacy: functional, communicative, and critical health literacy [19]. It was developed specifically for patients with type 2 diabetes but has subsequently been adapted for other populations [36]. It is based on a theoretical model developed by Nutbeam [37]—see Box 1 for definitions. Respondents rate their functional (five items), communicative (five items) and critical (four items) health literacy abilities using 4-point scales [19]. A numerical score is obtained for each scale by averaging the item
scores (Table A1 [available on the journal’s Web site at www.jclinepi.com]).

3.1.2.3. **eHealth Literacy Scale.** The eHealth Literacy Scale (eHEALS) was developed in Canada and measures an individual’s ability to use electronic health information for health care problems or issues in the broader population [18]. It is based on a model of eHealth literacy derived from social cognitive and self-efficacy theory (Table A3) [38]. Respondents are asked to indicate how strongly they agree or disagree with eight statements concerning their knowledge, confidence, and perceived skills relating to the use of electronic health information. It is unclear from the published literature how the eHEALS is scored and interpreted.

3.1.3. **Proxy measures of health literacy in the population**

3.1.3.1. **The demographic assessment of health literacy.** The demographic assessment of health literacy (DAHL) is a proxy indicator for health literacy based on the socio-demographic variables of age, sex, race/ethnicity, and years of schooling (Table A2 [available on the journal’s Web site at www.jclinepi.com]) [20]. It was derived from a large US population-based study (individuals older than 65 years) based on the S-TOFHLA. The development of the algorithm used associations between S-TOFHLA and the four demographic variables to generate equations (scoring weights), which are then applied to surveys to produce an imputed health literacy score and estimate the presence of limited health literacy at a population level.

3.1.3.2. **National Assessment of Adult Literacy.** The 2003 National Assessment of Adult Literacy (NAAL), a national literacy survey undertaken in the United States, included 28 items to measure health literacy in the population (Table A2 [available on the journal’s Web site at www.jclinepi.com]) [21]. These items address three health areas—clinical (three items), prevention (14 items), and navigation of the health care system (11 items)—and were developed to fit into one of the NAAL’s prose, document, and quantitative scales (see Box 2 for definitions) [21]. Items are not publicly available [21,39]. Each health literacy item is mapped to the relevant prose, document, or quantitative scale, and categorized as either below basic, basic, intermediate, or proficient (Table A2 [available on the journal’s Web site at www.jclinepi.com]) [21].

3.1.3.3. **Health Activities Literacy Scale.** The Health Activities Literacy Scale (HALS) was also developed in the United States to measure health literacy skills among adults in the population [22]. It consists of 191 health-related items, taken from large-scale adult literacy surveys conducted before 2003 and coded to represent health activities across five contexts or domains (Table A2 [available on the journal’s Web site at www.jclinepi.com]). All items relate to prose, document, or quantitative literacy. The individual
The strongest psychometric properties are unknown [40], and scoring properties for the HALS have not been specified [23].

3.1.3.4. Adult Literacy and Life Skills Survey. The Adult Literacy and Life Skills Survey (ALLS) includes a domain that measures health literacy in adult populations and has been administered in Canada and Australia [23,41]. It provides information on knowledge and skills across four areas: prose literacy, document literacy, numeracy, and problem solving. ALLS items appear to be a derivative of the HALS [23]. Proficiency of health literacy is measured on a 0–500 scale and grouped into five skill levels (Table A2 [available on the journal’s Web site at www.jclinepi.com]) [41]. Items of this survey are also unavailable.

3.2. Critical appraisal of psychometric properties

A summary of the psychometric attributes for each instrument is outlined in Tables 2, A4, and A5 and discussed in the following sections.

3.2.1. Content validity

Content validity evolves out of the planning and construction of an instrument and looks at the extent to which the items making up the instrument reflect the concept that is intended to be measured [42]. The underlying conceptual framework of health literacy varied across instruments. The TOFHLA and its derivatives were based on a refined definition of functional health literacy (Table 2) [9]; the content of questions in the SBSQ was based on five domains identified in a previous qualitative study [17,35]; and both FCCHL and eHEALS were based on specified theoretical models of health literacy and eHealth literacy, respectively (Table A4 [available on the journal’s Web site at www.jclinepi.com]) [37,38]. No clear underlying conceptual framework of health literacy was specified for the REALM, SAHLSA, NVS, or MART. The health literacy items selected for the NAAL were based on the definition of health literacy adopted in the US policy document “Healthy People 2010” (Table A5 [available on the journal’s Web site at www.jclinepi.com]). As acknowledged by the authors, although this definition implies broader skills, such as knowledge, items were limited to prose, document, and quantitative literacy in keeping with the overall literacy focus [21].

For most of the indices, there was a good description of how items were generated. The item content for the REALM and TOFHLA were developed from commonly used patient materials in the health care setting, whereas the NVS used a generic ice cream nutritional label that was chosen from four other scenarios on the basis of having the strongest psychometric properties [16].

Based on the underlying constructs or definitions specified by the authors, some measures did not cover all relevant intended content. For example, the construct of the TOFHLA is stated to be reading, writing, and numeracy skills in relation to health care, yet it does not include writing abilities. Similarly, although the SBSQ was based on five domains, the final SBSQ consists of only three items, and the eight items of the eHEALS do not appear to fully represent the six different types of literacy it purports to measure. Table A3 (available on the journal’s Web site at www.jclinepi.com) presents the constructs and items for SBSQ and eHEALS.

For population proxy indices, the DAHL presents as having good content validity, given it uses a credible range of demographic variables to derive a measure of health literacy. However, the weights across variables were derived from one sample of people aged 65 years or older, and this may not be generalizable to other ages and/or other settings. The content validity was unable to be evaluated for the other population surveys, as items are not publicly available [39].

For derivative instruments, how items were retained or excluded was well described and was based on psychometric estimates of item difficulty or statistical analyses.

3.2.2. Face validity

Face validity refers to whether an instrument appears to test what it is supposed to and that it is a plausible method for doing so [42]. Although the formats of each instrument appear plausible for their stated purpose, the face validity of most indices was questionable because of the difficulty in interpreting their scores. Most instruments collapse numerical scores into one of several categories but do not provide specific definitions for those categories. For some instruments, it was not possible to determine whether the categories were independent or mutually exclusive.

The REALM and its derivatives provide the most detailed definitions of their scoring categories. Categories correspond to school grade equivalents (Table 1) and have accompanying descriptions as to how simple the material or instructions should be to cater to different levels [8]. Conversely, the TOFHLA and its derivatives classify an individual as having either inadequate, marginal, or adequate health literacy, but provide no operational definitions of these categories. Instead, they identify tasks contained within the TOFHLA that individuals with suboptimal health literacy might have difficulty with. For example, patients with inadequate functional health literacy are expected to often misread medication dosing instructions, whereas those with marginal literacy are expected to have difficulty comprehending the Medicaid passage [34]. These examples are limited and may not be applicable to all individuals in these categories.

The SAHLSA used the TOFHLA to determine the cutoff scores between inadequate and adequate health literacy, but no descriptions for these categories are provided. The NVS scoring categories refer to the likelihood of inadequate or marginal literacy, but without clear accompanying definitions, these classifications appear to overlap. The FCCHL produces numerical scores for each level of health literacy.
No definitions are provided for scores; hence, it is unclear what constitutes appropriate health literacy levels.

For population surveys, scoring categories relate to “proficiency” of skills. Although definitions are outlined, these tend to be generic in terms of whether individuals have skills to be able to carry out tasks in everyday life rather than specific details relating to health-related activities.

The face validity of the HALS and MART could not be assessed. The HALS scoring algorithm is yet to be defined, and although the MART scoring system is based on the WRAT, specific details of categories and classifications were not outlined.

3.2.3. Construct validity

Construct validity refers to the degree to which an instrument measures what it is supposed to measure and can be assessed in various ways through factor analysis, hypothesis testing, and examining associations an instrument can be assessed in various ways through factor analysis, hypothesis testing, and examining associations an instrument can be assessed in various ways through factor analysis, hypothesis testing, and examining associations an instrument can be assessed in various ways through factor analysis, hypothesis testing, and examining associations an instrument can be assessed in various ways through factor analysis, hypothesis testing, and examining associations an instrument can be assessed in various ways through factor analysis.

The construct validity of health literacy indices that directly test an individual’s abilities has predominantly been assessed by comparing them with standardized reading tests and other health literacy instruments [8,9,15,16,24–27, 30,44–48] (Table 2). Both the REALM-S and TOFHLA have demonstrated strong correlation with the WRAT in adult populations in the United States—Pearson correlation coefficient = 0.88 and Spearman rank correlation = 0.74, respectively [9,24]. However, lower correlations with the WRAT were found for both measures in an adolescent population in the United States (Pearson correlation coefficients—REALM: 0.74 and TOFHLA [reading comprehension domain]: 0.6) [45]. The same measures have also been shown to weakly correlate with the Test of Reading Comprehension, an Australian reading assessment used with students in years 3–10 (Pearson correlation coefficients—TOFHLA: 0.39 and REALM: 0.36) [44,49]. The REALM has also correlated highly with the Slosson Oral Reading Test Revised (SORT-R); Peabody Achievement Test Revised; and the Basic Skills Assessment Initial Test, a measure of general literacy used in the United Kingdom [46] (Table 2). Derivatives of the REALM have also demonstrated high correlation with the WRAT [25,26]. The MART has not been compared with other literacy tests, and the SAHLSA used factor analysis (see later) to demonstrate construct validity.

Comparison of “direct testing” health literacy indices with one another has yielded variable results across studies. Both the TOFHLA and the reading comprehension domain of the S-TOFHLA have been shown to correlate highly with the REALM (Spearman rank correlation: 0.84 and 0.81, respectively) in the original studies [9,30]; however, subsequent studies have reported low to moderate correlations between the REALM and TOFHLA [44,45,47]. For example, the correlation between the reading comprehension section of the TOFHLA and REALM was moderate (Pearson correlation coefficient: 0.6) in an adolescent population study in the United States [45] and low in a study of patients attending a community-based rheumatology practice in Australia (Pearson correlation coefficient: 0.30) [44] and another Australian population-based study (Spearman rank correlation: 0.31) [47].

The NVS has been shown to correlate moderately with the TOFHLA in the US primary care study (Pearson correlation coefficient: 0.59) [16] and an Australian population-based study (Spearman rank correlation: 0.49) [47]. The NVS has also been shown to correlate moderately with the S-TOFHLA (Pearson correlation coefficient: 0.61) [48]. In contrast, the NVS has demonstrated low correlation with the REALM (Pearson correlation coefficient of 0.41 and Spearman rank correlation of 0.24, respectively) [47,48].

The SBSQ, SILS, and DAHL used the S-TOFHLA as the standard to assess the specificity and sensitivity of screening questions or model in the case of the DAHL (Tables A4 and A5 [available on the journal’s Web site at www.jclinepi.com]). For the three selected questions of the SBSQ, the optimum sensitivity and specificity appeared to be for the response “sometimes/somewhat” [17]. Subsequent testing of these questions with the REALM and S-TOFHLA suggests one possible question—“confidence with filling out forms” for detecting inadequate health literacy (area under the receiver operating characteristic curve = 0.74 [95% confidence interval (CI): 0.69–0.79] and 0.84 [95% CI: 0.79–0.89], respectively) [50]. For the SILS, a score ≥2 had 54% sensitivity and 83% specificity for detecting inadequate health literacy [28]. A cutoff score less than 62 in the DAHL resulted in 79% cases positively classified with a low sensitivity of 58% and an acceptable specificity of 84% [20].

Confirmatory factor analysis was undertaken to demonstrate construct validity for the REALM-SF and SAHLSA. This analysis identifies whether distinct underlying constructs or factors are identifiable in the data. Standardized fit statistics, for example, root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), and comparative fit index (CFI), provide evidence of whether distinct hypothesized constructs are present (see Ref. [51]). Both the REALM-SF (SRMR = 0.01, CFI = 0.97) and SAHLSA (RMSEA = 0.04, CFI = 0.85) have been shown to have good model fit (Table 2) [15,27].

For both the FCCHL and HHLT, demonstration of construct validity was limited, because other health literacy measures have not been translated into relevant languages. For the HHLT, scores were moderately correlated with patient self-reported reading ability (Pearson correlation coefficient: 0.67) [29].

3.2.4. Reliability

Reliability refers to the extent to which an instrument is consistent and free from error [42]. There are several approaches to reliability: establishing an instrument is capable of obtaining consistent results on separate occasions from
individuals in a stable environment (test–retest reliability) [52], assessing consistency of the results obtained by the same rater (intrarater reliability) and the agreement between the results of different raters (interrater reliability) [53]. Reliability coefficients generally range between 0.0 and 1.0, where 0.0 indicates that all measurement variation is attributable to error, and 1.0 indicates that the measurement has no error.

Five of the 19 indices reviewed reported a form of reliability, with test–retest reliability being most commonly cited. The REALM, REALM-S, and REALM-Teen demonstrate high test–retest reliability (reliability coefficients: 0.98, 0.99, and 0.98, respectively) when administered twice, 1 week apart (Table 2). The REALM was administered to 38 prison inmates and 26 substance abuse halfway house residents [8]. The REALM-S was administered to patients in primary care clinics and REALM-Teen to adolescents from high schools, primary care clinics, and summer programs in the United States, although numbers of individuals were not specified [24,25].

Test–retest reliability was also high for the SAHLSA (Pearson correlation coefficient: 0.86). It was administered twice to 40 Spanish-speaking patients in outpatient clinics, 2 weeks apart [15,21]. The test–retest reliability of the eHEALS was assessed by administering it to adolescents and young adults in the control arm of a combined randomized trial evaluation of a eHealth literacy promotion intervention and a Web-based smoking cessation program at four time points: pre- and postintervention (time points not specified) and at 3- and 6-month follow up [18]. Test–retest reliability was calculated between scores at each interval showing modest stability from baseline to 6 months ($r = 0.60–0.40$). Reported intraclass correlation coefficient of 0.49 suggests poor stability over time [18].

There have been no published studies of intrarater reliability for indices requiring interviewer administration, and there is only one study of interrater reliability. The REALM was shown to have high interrater reliability (Pearson correlation coefficient: 0.99) when administered by five research assistants to 20 patients at a university clinic [8].

Internal consistency is commonly measured by coefficient alpha and is another method for determining the degree to which items in a scale consistently measure a single construct [42]. Of the nine indices that reported internal consistency, the numeracy domain of the S-TOFHLA ($\alpha = 0.68$) and the critical health literacy domain of the FCCHL ($\alpha = 0.65$) had coefficients lower than the commonly accepted minimum $\alpha < 0.7$ [54] (Table 2).

### 3.2.6. Feasibility

This relates to how an instrument is administered and the requirements associated with the delivery. All indices that directly test individual abilities require an interviewer to administer all or some aspects of the test, although they appear easy to perform. Some interviewer training is required, particularly for the administration of the REALM and derivatives, SAHLSA and MART, which rely on subjective assessment of correct pronunciation. The estimated time to administer instruments was generally not detailed for self-reported and population measures (Tables A1 and A2 [available on the journal’s Web site at www.jclinepi.com]). Of those that were administered (Table 1), the 22 minutes to complete the TOFHLA could be considered a limitation.

### 3.2.7. Generalizability

This refers to whether the instrument can be effectively applied in different populations and settings. The REALM-S, TOFHLA, and S-TOFHLA have been used extensively in clinical populations in the United States. They have also been administered to clinical and general populations in the United Kingdom [46,56–59], Australia [44,47], and Brazil [60] (Table 2). Investigators have found it necessary to tailor the content of the TOFHLA to local settings in the United Kingdom and Australia but not the REALM or NVS [47,57].

### 4. Discussion

We have critically appraised the development and psychometric properties of 19 health literacy indices. Three approaches for measuring health literacy were identified: direct testing of an individual’s abilities, self-report of abilities, and population-based proxy measures. The underlying constructs assessed varied widely across instruments and most were not based on a specific conceptual framework. The content focused primarily on reading, comprehension, and numeracy skills; scoring categories were poorly defined and may not be mutually exclusive; only five had been assessed for reliability. Responsiveness to change has not been assessed for any of the indices. Overall, the TOFHLA and REALM-S had the strongest psychometric properties.

The variability in the measurement approach across instruments reflects the view that health literacy is a complex and multifaceted concept that is still evolving [61,62]. Health literacy has developed through a convergence of two main areas of study: (1) health education and promotion where health literacy is considered a personal “ asset” and (2) clinical care where it is considered a “risk” that needs to be managed to achieve positive health outcomes [40,63]. Such context may account for differences in measurement approaches. For example, indices, such as the FCCHL and eHEALS, focus more on personal abilities,
whereas the TOFHLA and SBSQ look to identify individuals at risk.

However, the lack of an explicit definition of the concept that many health literacy indices were developed to measure limited our ability to make fully informed judgments about their face and content validities. Based on a widely accepted definition of health literacy [40, 64], none of the indices we evaluated seemed to fully measure a person’s ability to seek, understand, and use health information. This calls into question their overall relevance and usefulness.

To tailor health information to a patient’s needs, a health literacy instrument should ideally be capable of clearly describing an individual patient’s abilities and areas that need to be addressed. Without a clear understanding of what different categories within a health literacy measure mean, it is difficult to know how to apply this information in a clinical or public health context. At a population level, inability to clearly discriminate between groups of people with different levels of health literacy may hamper efforts to target interventions toward “high-risk” groups.

In the absence of a “gold standard,” empirical evidence of construct validity of most health literacy measures was obtained by comparison with standardized literacy tests and/or other health literacy indices. Several health literacy instruments were developed from standardized literacy tests, and therefore, it is not surprising that there was a strong correlation between them. However, there was not always a strong correlation between health literacy measures, implying that they might have measured different constructs. None of the instruments we assessed have been compared with broader elements of health literacy implied in definitions, such as communication and self-management skills [65, 66].

Other authors have also identified a mismatch between definitions and measurement of health literacy [28, 61, 62, 67] and the need for more comprehensive measures [64, 68]. In a qualitative study to conceptualize health literacy from the patient perspective, we identified a range of individual abilities and contextual factors important in seeking, understanding, and using health information [69]. Individual abilities included knowing when to seek health information, verbal communication, assertiveness, capacity to process and retain information, and applying information to lifestyle, which all reflect broader constructs not incorporated in existing instruments.

Although there were no studies that specifically assessed the responsiveness of health literacy instruments, one study was identified that examined changes in health literacy over time using the REALM-S [70]. In a randomized control trial that tested the effectiveness of a literacy learning program for individuals with both depression and low health literacy, health literacy scores in the intervention group increased by an average of 7 points between baseline and final follow-up. This suggests that the REALM-S is responsive to change; however, empirical demonstration of responsiveness over time is critical if instruments are to accurately measure the effectiveness of future programs to improve health literacy in populations.

A strength of our review is the use of a structured methodological framework to critically appraise existing health literacy measures. Based on our evaluation, the REALM and TOFHLA appear to have the strongest psychometric properties, although each has its limitation. Both state a clear purpose, have been compared with a range of indices to demonstrate construct validity, are simple to understand and administer, and have been used in clinical and population settings. A limitation of our study was our inability to fully appraise the population measures because of limited information being available in the public domain.

Our study indicates that further research is needed to address psychometric gaps. For example, in-depth qualitative studies may assist in developing cutoff scores or categories that can be meaningfully applied to clinical or population settings [71]. Additionally, randomized control trials of program interventions to improve health literacy would be valuable forums to obtain information on responsiveness and reliability of existing measures. Future instruments need to encompass broader concepts of health literacy identified in existing literature [72]. This will require in-depth consultation with key stakeholders to clearly articulate a conceptual basis of health literacy and identify clear constructs to measure.

Acknowledgments

Richard H. Osborne is supported in part by a National Health and Medical Research Council (NHMRC) Population Health Fellowship (Career Development Award), and Rachelle Buchbinder is supported in part by an NHMRC Practitioner Fellowship. Joanne E. Jordan is supported by an Australian Postgraduate Award.

Appendix

Supplementary material

Supplementary material can be found, in the online version, at 10.1016/j.jclinepi.2010.04.005

References


